DOCUMENT RESUME

ED 471 081 TM 034 586

AUTHOR	Miller, M. David
TITLE	Generalizability of Performance-Based Assessments. Technical Guidelines for Performance Assessment.
INSTITUTION	Council of Chief State School Officers, Washington, DC.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	2002-00-00
NOTE	37p.; Produced by the State Collaborative on Assessment and Student Standards.
CONTRACT	R279A50006
AVAILABLE FROM	Council of Chief State School Officers, One Massachusetts Avenue, Suite 700, Washington, DC 20001-1431 ($10). Tel: 202-408-5505; Fax: 202-408-8072. For full text: http://publications.ccsso.org
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01 Plus Postage. PC Not Available from EDRS
DESCRIPTORS	Elementary Secondary Education; Error of Measurement; *Generalizability Theory; *Performance Based Assessment; *Reliability; *State Programs; Testing Programs

ABSTRACT

In 1994 the State Collaborative on Assessment and Student Standards of the Council of Chief State School Officers began a study to examine the generalizability of performance-based assessments (PBAs) for state-mandated assessment programs. The intent was to examine the major sources of error associated with PBAs and the generalizability and dependability of the assessments. Also examined were the consistency of results of generalizability studies across content areas and the generalizability of student- and school-level data. Generalizability analyses were conducted using state performance-based assessments from four states: Alabama, Connecticut, Delaware, and North Carolina. Most of the assessment programs are using PBAs that have a reasonable level of generalizability (above 0.70) at the student level. However, the level of generalizability is not as high as that found for more traditional forms of assessment. Other advantages of PBAs, such as their consequences for assessment, are necessary to justify their use. Results at the school level are mixed. While the standard errors are smaller with reasonable school sizes, allowing the use of school means with confidence intervals, the lower variance component sometimes associated with schools results in low values for generalizability coefficients and the dependability indices. Although limited data are available, it appears from the Connecticut assessment that tests with more extended responses need fewer items to achieve the same level of generalizability. Even with extended responses, at least two tasks are needed to achieve reasonable levels of generalizability. (Contains 43 tables and 5 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

ED 471 081

SCASS
State Collaborative on Assessment and Student Standards

Generalizability of Performance-Based Assessments

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

B. Buterbaugh

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM034586

M. David Miller

SCASS TGPA
Technical Guidelines for Performance Assessment

A Publication of the Council of Chief State School Officers

BEST COPY AVAILABLE

The Council of Chief State School Officers (CCSSO) is a nationwide, nonprofit organization composed of the public officials who head departments of elementary and secondary education in states, the District of Columbia, the Department of Defense Education Activity, and five extra-state jurisdictions. CCSSO seeks its members' consensus on major education issues and expresses their views to civic and professional organizations, to federal agencies, to Congress, and to the public. Through its structure of standing committees and special task forces, the Council responds to a broad range of concerns about education and provides leadership on major education issues.

Because the Council represents each state's chief education administrator, it has access to the educational and governmental establishment in each state and to the national influence that accompanies this unique position. CCSSO forms coalitions with many other education organizations and is able to provide leadership for a variety of policy concerns that affect elementary and secondary education. Thus, CCSSO members are able to act cooperatively on matters vital to the education of America's young people.

The State Education Assessment Center was established through a resolution by the membership of CCSSO in 1984. This report is sponsored by the Assessment Center's State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TGPA) consortium. The SCASS TGPA works with researchers to design and implement practical and timely research on large-scale performance assessment. This research provides information useful in designing state assessment and accountability programs so that they yield results that can be used to improve student learning.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS

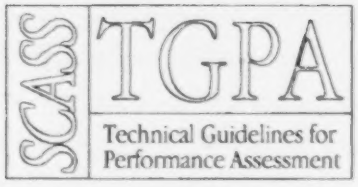
G. Thomas Houlihan
Executive Director

Wayne N. Martin
Director
State Education Assessment Center

John F. Olson
Director of Assessments
State Education Assessment Center

Phoebe C. Winter
Project Director, Technical Guidelines for Performance Assessment
State Collaborative on Assessment and Student Standards

© Copyright 2002 by the Council of Chief State School Officers, Washington, DC. This document may not be reproduced without the permission of the Council of Chief State School Officers, the copyright holder.



Generalizability of Performance-Based Assessments

M. David Miller
University of Florida

Council of Chief State School Officers
One Massachusetts Avenue, N.W.
Washington, DC 20001
202-408-5505

This report was prepared for submission under contract with the Council of Chief State School Officers and funded by the U.S. Department of Education, Office of Educational Research and Improvement, Grant No. R279A50006. The views and opinions expressed in this report are not necessarily those of the U.S. Department of Education, the Council of Chief State School Officers, the State Collaborative on Assessment and Student Standards, or the states participating in the study.

Acknowledgments

This research project was conducted with the assistance of the following members of the State Collaborative on Assessment and Student Standards, Technical Guidelines for Performance Assessment, who helped develop the research design, assisted in the data collection, and reviewed draft versions of the report.

STUDY GROUP ON THE GENERALIZABILITY OF PERFORMANCE-BASED ASSESSMENTS

Peter Behuniak	Connecticut Department of Education, Chair
Gloria Turner	Alabama Department of Education
Liru Zhang	Delaware Department of Education
Jonathan Dings	Kentucky Department of Education*
Ellie Sanford	North Carolina Department of Public Instruction*
Ellen Hedlund	Rhode Island Department of Education
Gordon Ensign	Washington State Commission on Student Learning*
Duncan MacQuarrie	Washington Department of Public Instruction*
Robert Linn	University of Colorado
Doris Redfield	Consultant to the Council of Chief State School Officers*
John Olson	Council of Chief State School Officers
Edward Roeber	Council of Chief State School Officers*
Phoebe Winter	Council of Chief State School Officers*

* At time of development.

Executive Summary

In 1994, the State Collaborative on Assessment and Student Standards (SCASS), Technical Guidelines for Performance Assessment (TPGA) began a study to examine the generalizability of performance-based assessments (PBAs) for state-mandated assessment programs. The intent was to examine (a) the major sources of error associated with PBAs, as well as the generalizability/dependability of the assessments, (b) the consistency of the results of the generalizability studies across content areas and grade levels, and (c) the generalizability of student- and school-level data.

Generalizability analyses were conducted using state performance-based assessments. Data were presented from four states: Alabama, Connecticut, Delaware, and North Carolina. Generalizability was examined at the elementary through high school levels. The assessments represented a broad range of content areas, including mathematics, reading, writing, social studies, literature, and interdisciplinary studies. Generalizability coefficients, dependability indices, and standard errors were reported at the student and school levels. On the basis of the data analyses, the following conclusions and recommendations were reached:

- Most assessment programs are using PBAs that have a reasonable level of generalizability (above .70) at the student level. However, the level of generalizability is not as high as that found for more traditional forms of assessment. Thus, other advantages of PBAs—such as their consequences for instruction—are necessary to justify their use.

- Results at the school level are mixed. While the standard errors are smaller with reasonable school sizes, allowing the use of school means with confidence intervals, the lower variance component sometimes associated with schools results in low values for the generalizability coefficients and the dependability indices. This may be a function of the sampling designs in Connecticut (pilot) and North Carolina (field test). Even the findings in grade 8 in Delaware, where a statewide assessment was used, may be a function of using only larger schools to estimate the variance components.

- The relative magnitude of the variance components and the largest sources of error are consistent with the findings of Brennan (1996), suggesting that number of tasks has a greater effect than number of raters.

- Results were better than those previously reported by Shavelson, Baxter, and Gao (1993). However, the lower number of tasks needed may be the result of multiple factors, including how narrowly the task domain is defined and the effects of high-stakes testing on teaching to the test. It should also be noted that Shavelson et al. based their conclusions on a minimum level of generalizability equal to .80. (Even at .80, these results suggest fewer tasks are needed).

- Although limited data are available, it appears from the Connecticut assessments that tests with more extended responses need fewer items to achieve the same level of generalizability. However, even with extended responses, at least two tasks are needed to achieve reasonable levels of generalizability.

Contents

INTRODUCTION 1

Table 1: Characteristics Of Participating States' Assessment Programs 1

OVERVIEW OF PARTICIPATING STATES' ASSESSMENT PROGRAMS 2

Alabama 2
Connecticut 2
Delaware 2
North Carolina 3
Uses of Assessment Results 4

RESULTS OF THE GENERALIZABILITY STUDY 5

Alabama 5
Table 2: Variance Components—Alabama 5
Student-Level Analyses 5
Table 3: Generalizability Coefficients—Alabama 6
Table 4: Dependability Indices—Alabama 6
Table 5: Standard Errors for Relative Decisions—Alabama 6
Table 6: Standard Errors for Absolute Decisions—Alabama 6
School-Level Analyses 6
Table 7: Generalizability Coefficients for Schools—Alabama 7
Table 8: Dependability Indices for Schools—Alabama 7
Table 9: Standard Errors for Relative Decisions (Schools)—Alabama 8
Table 10: Standard Errors for Absolute Decisions (Schools)—Alabama 8
Connecticut 8
Table 11: Variance Components—Connecticut 8
Student-Level Analyses 9
Table 12: Generalizability Coefficients—Connecticut 9
Table 13: Dependability Indices—Connecticut 9
Table 14: Standard Errors for Relative Decisions—Connecticut 9
Table 15: Standard Errors for Absolute Decisions—Connecticut 10
School-Level Analyses 10
Table 16: Generalizabilty Coefficients (Schools)—Connecticut 10
Table 17: Dependability Indices (Schools)—Connecticut 11
Table 18: Standard Errors for Relative Decisions (Schools)—Connecticut 12
Table 19: Standard Errors for Absolute Decisions (Schools)—Connecticut 13
Delaware 13
Table 20: Variance Components—Delaware 13
Student-Level Analyses 14
Table 21: Generalizabilty Coefficients—Delaware 14
Table 22: Dependability Indices—Delaware 14
Table 23: Standard Errors for Relative Decisions—Delaware 14
Table 24: Standard Errors for Absolute Decisions—Delaware 15
Table 25: Variance Components—Delaware 15
Table 26: Generalizability Coefficients/Dependability Indices—Delaware 15
Table 27: Standard Errors—Delaware 15
School-Level Analyses 15
Table 28: Generalizability Coefficients (Schools)—Delaware 16
Table 29: Dependability Indices (Schools)—Delaware 17
Table 30: Standard Errors For Relative Decisions (Schools)—Delaware 18
Table 31: Standard Errors for Absolute Decisions (Schools)—Delaware 19
Table 32: Generalizability Coefficients/Dependability Indices (School)—Delaware 19
Table 33: Standard Errors (School)—Delaware 20

North Carolina 20
Table 34: Variance Components—North Carolina *21*
Student-Level Analyses 21
Table 35: Generalizability Coefficients—North Carolina *21*
Table 36: Dependability Indices—North Carolina *22*
Table 37: Standard Errors for Relative Decisions—North Carolina *22*
Table 38: Standard Errors for Absolute Decisions—North Carolina *23*
Table 39: Variance Components—North Carolina *23*
Table 40: Generalizability Coefficients/Dependability Indices—North Carolina *23*
Table 41: Standard Errors –North Carolina *23*
School-Level Analyses 23
Table 42: Generalizability Coefficients/Dependability Indices (School)—North Carolina *24*
Table 43: Standard Errors (School)—North Carolina *24*

CONCLUSIONS AND RECOMMENDATIONS 25

REFERENCES 26

APPENDIX 27

FORMULAS USED IN GENERALIZABILITY ANALYSES 27

Alabama – Student Level 27
Alabama – School Level 27
Connecticut – Student Level 28
Connecticut – School Level 28
Delaware – Reading, Math 28
Delaware – Student Level (Writing) 28
Delaware – School Level (Writing) 29
North Carolina – Student Level (Math, Social Studies, and Reading) 29
North Carolina – English 29

Introduction

In 1994, the Technical Guidelines for Performance Assessment (TPGA) State Collaborative on Assessment and Student Standards (SCASS) began a study to examine the generalizability of performance-based assessments (PBAs) for state-mandated assessment programs. The intent was to examine (a) the major sources of error associated with PBAs, as well as the generalizability/dependability of the assessments, (b) the consistency of the results of the generalizability studies across content areas and grade levels, and (c) the generalizability of student- and school-level data.

This study was conducted because there is no large body of literature suggesting the answers to these issues. Brennan (1996) summarized what was known about PBAs in a technical report for the National Center for Education Statistics. He cited six data sets that had analyzed the generalizability of PBAs: a voluntary science assessment from the California Assessment Program (see Shavelson, Baxter and Gao, 1993), a science and math assessment for a local education agency (see Shavelson et al., 1993), a performance task for the Multistate Bar Examination (see Gamache and Brennan, 1994), and listening and writing tests for a new program (Work Keys) developed by ACT (see Brennan, Gao, & Colton, 1995). Brennan found the results of the six PBAs to be fairly similar, in that the person-by-task variance component was generally the largest, indicating "only a limited degree of across-task generalizability" (p. 27). On the other hand, the rater, the person-by-rater, and the task-by-rater variance components were typically small in magnitude, indicating high generalizability across raters. Similar data were also reported at the group level by Gao, Brennan, and Shavelson (1994), based on data from the California Assessment Program.

For this study, four states—Alabama, Connecticut, Delaware, and North Carolina—volunteered the use of their PBA data. The characteristics of the PBA programs are summarized in Table 1. This report is based on the analyses of data from 40 performance-based assessments in the four states collected in 1995. These assessments vary widely in several areas. They are given in grades 3 through high school and cover content in mathematics, reading, writing, social studies, literature, and interdisciplinary studies. The PBA tasks varied from short-answer items to tasks requiring students to produce extended responses. In addition, the data collection varied from statewide administrations to smaller pilot studies with assessments that were not always operationally administered. The sources of error include various combinations of students, raters, and tasks. The universes of generalization include the student and the school.

TABLE 1: CHARACTERISTICS OF PARTICIPATING STATES' ASSESSMENT PROGRAMS

State	Grade Levels	Content Areas	Form of Data
Alabama	High School	Algebra, Geometry	Statewide
Connecticut	High School	Literature, Interdisciplinary	Pilot
Delaware	3,5,8,10	Reading, Writing, Mathematics	Statewide
North Carolina	3,4,5,6,7,8, High School	Mathematics, Reading, Social Studies, English	Field Test

The format of this report is as follows: (1) descriptions of the state testing programs, (2) presentations of the data by state (in alphabetical order), and (3) a discussion of the overall conclusions and implications.

Overview of Participating States' Assessment Programs[1]

ALABAMA

Data from the high school algebra and geometry PBAs were used for this study. These assessments were given to all students completing courses in the respective content area. The algebra assessment contains both constructed-response (35%) and multiple-choice and grid-in (65%) items; 53% of the geometry test is multiple-choice and 47% is constructed-response. The purposes of the assessments are to provide public accountability, to monitor the implementation of the curriculum, and to provide information that can be used for program improvement. The algebra assessment program was implemented in spring 1992 and geometry in spring 1994. In addition, Alabama has more traditional norm-referenced assessments, basic competency assessments, and direct writing assessments. The more traditional assessment results are used as exit exams for graduation and for school accountability, as well as for monitoring student achievement.

CONNECTICUT

For this study, data were used from the high school Connecticut Academic Performance Test (CAPT), which is required of all students. The CAPT assessments cover mathematics, science, responding to literature, editing, and an interdisciplinary section. Multiple-choice, short constructed-response, and extended written-response items and performance activities are used on the test, with the proportion of each type varying by content area.

Both the interdisciplinary task and the responding-to-literature component are extended-response activities. For the interdisciplinary component, students have 90 minutes to read the source materials and complete their response. The student response is scored holistically on a 6-point rubric by two or more raters. For the responding-to-literature component, students have 90 minutes to read the literature and respond to six questions about the material. Student responses to all six questions are scored holistically, considering all of the student's responses to the entire set of questions. Two or more raters employ a 6-point holistic rubric to generate a simple overall score. The purposes of the test are to monitor student achievement, to award Certificates of Mastery, and to identify program improvement needs. The tests were implemented in 1995. In addition, Connecticut has assessments in mathematics, reading, and writing in grades 4, 6, and 8. These results are used for monitoring student achievement, individual instruction, program improvement, allocating funding, and curriculum development.

DELAWARE

The assessments used in this study were part of the assessment system in place in 1993 to 1995. For this study, data were used from all grade levels taking the PBAs.

[1] See state web sites for current and complete information about states' testing programs.

The assessments covered mathematics, reading, and writing in grades 3, 5, 8, and 10. The purposes of the tests were to monitor student achievement and to improve teaching and learning. In addition, Delaware has norm-referenced assessments in grades 3, 5, 8, and 10 that are used for linking assessments over time. Test results are used for monitoring student achievement, curriculum development, and improving teaching and learning.

NORTH CAROLINA

North Carolina offers a comprehensive student assessment system, which was redesigned in 1996.[2] For this study, data were used from field-test PBAs in grades 3 through high school. The assessments used in the study cover reading, mathematics, and social studies in grades 3 to 8; writing in grades 4, 6, and 8; and algebra, geometry, and English in high school. All items on the assessments included in this study were constructed-response.

The North Carolina Open-Ended Tests were designed to measure broad higher-level thinking skills by requiring students to apply or demonstrate skills and knowledge beyond the recall level. These items were designed to be open so that the quality of the student's responses would determine his or her score. Each form of the test contained items that assessed (1) the reading strand of the English Language Arts Standard Course of Study, (2) the mathematics Standard Course of Study, and (3) the social studies Standard Course of Study. Some of the mathematics items required the production of a specific answer to a problem, but the student was also asked to explain how he or she arrived at the answer. The explanation helped to determine the student's score.

Three forms of the assessments studied were administered in each classroom. Each form consisted of 3 or 4 reading items, 3 or 4 mathematics items, and 3 or 4 social studies items, for a total of 10 items on each form. Each item was scored on a 3- or 4-point scale and took students about 10 minutes to complete. Scores on the open-ended tests were reported at the school and school-system levels in terms of developmental scale scores and percentiles.

The North Carolina Test of English II was developed to measure student achievement in writing based on four basic composing criteria: main idea, supporting detail, organization, and coherence. The test used in the study was a two-item test administered across two days; administration time was 55 minutes per prompt. On the first day, all students responded to a core prompt that requires literary analysis focused on a literary concept in a work that is defined as World Literature (not British or American). On the second day, students responded to one of four prompts that assess the four modes of writing. Each prompt was scored using a 6-point focused holistic scale (1 to 6). A score of 0 indicated a misunderstanding of the meaning of the literary term that was being assessed. Scores on the core prompt were reported at the student, school, and system levels in terms of raw scores; scores on the variable prompt were reported at only the school and school-system levels in terms of raw scores.

The purposes of the PBAs are to give teachers more information about student achievement; multiple-choice assessments are used for school accountability. The majority of the assessments were first implemented in 1993; the exceptions were the

[2] For information about the current North Carolina assessment program, visit the North Carolina Department of Public Instruction's web site.

direct writing assessments in grades 6 and 8 (1987) and the high school assessments in algebra (1994), geometry (1989), and English (1991).

Uses of Assessment Results

The four states used results of the assessments in the study in a variety of ways. In Alabama, test scores are reported to the public at the school, district, and state levels. Schools receive rosters of individual student scores, but students do not receive individual score reports. Teachers use results for instructional planning, and school and district results are reported to the public.

Connecticut reports scores at the student, classroom, school, district, region, and state levels. Individual scores are reported to students and parents, and the results are used as part of retention and promotion decisions, and often for adjustments in student programming. Teachers use assessment results for instructional planning. Teacher evaluations are based in part on the test results, school and district results are reported to the public, and the statewide funding formula for districts includes test results as one factor.

Delaware reported the 1993–1995 assessment results at the student, school, district, and state levels, and school and district reports were made public. Student-level decisions were not based on test scores, but teachers used the results for instructional planning.

North Carolina reported assessment results at the school, system, and state levels, and school and system reports were made public. Teachers used the results for instructional planning and were given scoring guides and examples of student work to assist them. Students and parents also received reports of student scores on the English II assessment.

Results of the Generalizability Study

Variance estimates, generalizability and dependability coefficients, and standard errors for both relative and absolute decisions were calculated for various configurations of the test data, for both student and school results. Generalizability and dependability coefficients were compared against a criterion of .70. While the criterion varies across studies and assessments, .70 seems to be a lower bound of what would be acceptable. Other, more rigorous criteria could be examined with the data presented.

ALABAMA

The data for Alabama are based on a statewide assessment during spring, 1995 (similar findings were found for two preceding years and are not included in this report). The two assessment programs are algebra and geometry at the high school level (open-ended problems with a single rater). To examine the generalizability at the student and school levels, 10 schools were randomly drawn from a pool of schools with at least 50 students, and 50 students were randomly drawn within each school. The seven algebra items each had scores ranging from 0 to 3. The six geometry items varied in the points per item and were rescaled from 0 to 1 to provide a common scale (dividing scores by the maximum points allowed). The range of scores was somewhat restricted by the overall difficulty of the items. For example, the item means for the seven algebra items were .86, 1.37, .47, .96, .77, and 1.30 on a 3-point scale (similar trends were observed for the geometry items).

The variance components for the two assessments are reported in Table 2. As can be seen, the largest source of error is the task-by-pupil interaction (consistent with Brennan, 1996) and the variance component associated with pupils is relatively large. The differences between the two content areas represent the differences in the scales (algebra 0 to 3; geometry 0 to 1).

TABLE 2: VARIANCE COMPONENTS—ALABAMA

Test	Sc	Pu:Sc	Tsk	Tsk*Sc	Tsk*Pu:Sc,E
Algebra	0.050220	0.24700	0.10075	0.05674	0.721
Geometry	0.004193	0.02233	0.00378	0.00256	0.049

Sc=School; Pu=Pupil; Tsk=Task; E=Error

Student-Level Analyses

The generalizability of the assessments with 5 and 10 items, and the actual number of items used, are reported in Table 3. These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative (as in norm-referenced testing). As can be seen, both assessments have a generalizability coefficient in excess of .70 under the current measurement conditions.

TABLE 3: GENERALIZABILITY COEFFICIENTS—ALABAMA

Test	5 Items	Actual	10 Items
Algebra	0.65645	0.72790	0.79260
Geometry	0.72008	0.75532	0.83726

The dependability of the assessments with 5 and 10 items, and the actual number of items used, are reported in Table 4. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a student knows (as in criterion-referenced testing). As can be seen, both assessments have a dependability coefficient in excess of .70 under the current measurement conditions. As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 4: DEPENDABILITY INDICES—ALABAMA

Test	5 Items	Actual	10 Items
Algebra	0.62848	0.70312	0.77186
Geometry	0.70558	0.74199	0.82738

The standard errors for relative and absolute decisions are reported in Tables 5 and 6, respectively. These are the standard errors that could be used to make confidence intervals around the students' scores (e.g., a 95% confidence interval could be created by using the individual's score and 1.96 times the standard error). As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and would need to be transformed for use with scores other than the average item score.

TABLE 5: STANDARD ERRORS FOR RELATIVE DECISIONS—ALABAMA

Test	5 Items	Actual	10 Items
Algebra	0.39440	0.33333	0.27888
Geometry	0.10155	0.09270	0.07181

TABLE 6: STANDARD ERRORS FOR ABSOLUTE DECISIONS—ALABAMA

Test	5 Items	Actual	10 Items
Algebra	0.41916	0.35426	0.29639
Geometry	0.10521	0.09604	0.07439

School-Level Analyses

The generalizability of the assessments with 5 and 10 items, and the actual number of items used, are reported in Table 7 for examining school means based on 20, 50, and 80 students within each school. These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, both assessments have a generalizability coefficient in excess of .70 under the current measurement conditions when the school sizes are 50 or 80. However, generalizability does not exceed .70 with a school size of 20.

TABLE 7: GENERALIZABILITY COEFFICIENTS FOR SCHOOLS—ALABAMA

		N = 20	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.61902	0.66231	0.69897
Geometry	0.66434	0.68240	0.72162
		N = 50	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.72371	0.76876	0.80641
Geometry	0.78409	0.80178	0.83967
		N = 80	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.75567	0.80095	0.83864
Geometry	0.82110	0.83846	0.87548

The dependability of the assessments with 5 and 10 items, and the actual number of items used, are reported in Table 8 for school means based on 20, 50, or 80 students. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a school gets correct. As can be seen, the geometry assessment has a dependability coefficient in excess of .70 under the current measurement conditions with 50 or more students in a school. However, the dependability index does not exceed .70 for the algebra assessment, even with 80 students. As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 8: DEPENDABILITY INDICES FOR SCHOOLS—ALABAMA

		N = 20	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.49587	0.55665	0.61301
Geometry	0.59322	0.61888	0.67750
		N = 50	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.56086	0.62997	0.69412
Geometry	0.68689	0.71550	0.78053
		N = 80	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.57986	0.65142	0.71787
Geometry	0.71512	0.74457	0.81138

The standard errors for relative and absolute decisions are reported in Tables 9 and 10, respectively. These are the standard errors that could be used to make confidence intervals around the schools' mean scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and would need to be transformed for use with scores other than the average item score. Despite the lower dependability/generalizability coefficients for the schools, the standard errors are lower than expected. The poorer generalizability coefficients are in part due to the differences in the universe score variance associated with schools versus students, and the lower standard errors still provide a useful way of examining school means.

TABLE 9: STANDARD ERRORS FOR RELATIVE DECISIONS (SCHOOLS)—ALABAMA

		N = 20	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.17581	0.16002	0.14707
Geometry	0.04603	0.04418	0.04022
		N = 50	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.13846	0.12291	0.10980
Geometry	0.03398	0.03220	0.02830
		N = 80	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.12743	0.11172	0.098299
Geometry	0.03023	0.02842	0.024422

TABLE 10: STANDARD ERRORS FOR ABSOLUTE DECISIONS (SCHOOLS)—ALABAMA

		N = 20	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.22596	0.19999	0.17806
Geometry	0.05362	0.05082	0.04468
		N = 50	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.19830	0.17175	0.14876
Geometry	0.04372	0.04083	0.03434
		N = 80	
Test	**5 Items**	**Actual**	**10 Items**
Algebra	0.19075	0.16393	0.14049
Geometry	0.04087	0.03793	0.03122

CONNECTICUT

The data for Connecticut are based on a pilot assessment during spring 1995. (Note: Some of the assessments with the weaker results shown here were dropped before assessments were created for statewide usage.) The assessment programs are interdisciplinary (IN31, IN32, and IN33) and literature (LT21, LT22, LT23, and LT24) at the high school level (open-ended problems with two raters). The interdisciplinary assessments all had the same common task with a second task, as did the literature assessments. To examine the generalizability at the student and school levels, seven to nine schools were drawn from a pool of schools with 40 or more students, and 40 students were randomly drawn within each school. The variance components for the seven assessments are reported in Table 11. As can be seen, the largest source of error is the highest-order interaction (raters within other effects), and the variance component associated with pupils is relatively large.

TABLE 11: VARIANCE COMPONENTS—CONNECTICUT

Test	N	Sc	Pu:Sc	Tsk	Tsk*Sc	Tsk*Pu:Sc	Rat:PST,E
IN31	8	0.05003	0.46500	0.31599	0.053863	0.1720	0.320
IN32	7	0.02214	0.61375	0.01584	0.000863	0.1570	0.304
IN33	8	0.22018	0.41225	0.12663	0.014512	0.1870	0.447
LT21	9	0.21752	0.55900	0.17558	0.023887	0.1620	0.349
LT22	7	0.06174	0.67600	0.02805	0.000000	0.1220	0.319
LT23	8	0.10063	0.75450	0.01786	0.053963	0.2055	0.266
LT24	7	0.18293	0.74350	0.07831	0.020125	0.1475	0.247

Sc=School; Pu=Pupil; Tsk=Task; Rat=Rater; E=Error

Student-Level Analyses

The generalizability of the assessments with two and four items, and with one and two raters, are reported in Table 12. These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, all assessments have a generalizability coefficient in excess of .70 with two tasks and two raters. In fact, most have a generalizability coefficient in excess of .70 with two tasks and one rater.

TABLE 12: GENERALIZABILITY COEFFICIENTS—CONNECTICUT

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.65362	0.72748	0.79054	0.84225
IN32	0.73359	0.80409	0.84632	0.89141
IN33	0.66106	0.74849	0.79595	0.85616
LT21	0.74382	0.81165	0.85309	0.89604
LT22	0.76989	0.83978	0.86999	0.91292
LT23	0.76497	0.81336	0.86684	0.89707
LT24	0.81714	0.86421	0.89937	0.92716

The dependability of the assessments with two and four items, and with one and two raters, are reported in Table 13. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a student knows. As can be seen, all assessments except IN31 (and rounding error for IN33) have a dependability coefficient in excess of .70 with two raters and two tasks. As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 13: DEPENDABILITY INDICES—CONNECTICUT

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.54445	0.59475	0.70504	0.74589
IN32	0.72695	0.79612	0.84189	0.88649
IN33	0.62003	0.69631	0.76545	0.82097
LT21	0.68612	0.74343	0.81385	0.85284
LT22	0.75879	0.82659	0.86285	0.90506
LT23	0.75891	0.80650	0.86293	0.89289
LT24	0.78986	0.83376	0.88260	0.90934

The standard errors for relative and absolute decisions are reported in Tables 14 and 15, respectively. These are the standard errors that could be used to make confidence intervals around the students' scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations.

TABLE 14: STANDARD ERRORS FOR RELATIVE DECISIONS—CONNECTICUT

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.52243	0.43924	0.36941	0.31059
IN32	0.48055	0.39361	0.33980	0.27833
IN33	0.56944	0.46098	0.40265	0.32596
LT21	0.51715	0.42449	0.36568	0.30016
LT22	0.46957	0.37517	0.33204	0.26528
LT23	0.51257	0.44298	0.36244	0.31323
LT24	0.45532	0.38153	0.32196	0.26978

TABLE 15: STANDARD ERRORS FOR ABSOLUTE DECISIONS—CONNECTICUT

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.65645	0.59239	0.46418	0.41888
IN32	0.48872	0.40355	0.34558	0.28535
IN33	0.62255	0.52519	0.44021	0.37136
LT21	0.59601	0.51767	0.42145	0.36605
LT22	0.48428	0.39342	0.34244	0.27819
LT23	0.52121	0.45295	0.36855	0.32028
LT24	0.49646	0.42979	0.35105	0.30391

School-Level Analyses

The generalizability of the assessments with two and four items, and with one and two raters, are reported in Table 16 for examining school means based on 20, 50, and 80 students within each school. These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, most of the literature assessments have a generalizability coefficient in excess of .70 when the school sizes are 50 or 80. However, generalizability does not exceed .70 with a school size of 20 or for the larger school sizes with most of the interdisciplinary assessments. This is linked to the low variance component associated with schools on these assessments (the standard errors can still be used, even in the absence of universe score variance).

TABLE 16: GENERALIZABILTY COEFFICIENTS (SCHOOLS)—CONNECTICUT

N = 20

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.44467	0.46106	0.53857	0.55042
IN32	0.34173	0.36302	0.37647	0.38904
IN33	0.83434	0.85238	0.87253	0.88230
LT21	0.80507	0.81828	0.84366	0.85086
LT22	0.57938	0.60190	0.61098	0.62328
LT23	0.56814	0.57901	0.63795	0.64475
LT24	0.76211	0.77205	0.79512	0.80049

N = 50

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.54869	0.55849	0.66481	0.67195
IN32	0.56110	0.58358	0.59939	0.61199
IN33	0.90976	0.91823	0.93605	0.94052
LT21	0.88511	0.89144	0.91693	0.92031
LT22	0.77496	0.79079	0.79701	0.80531
LT23	0.68263	0.68884	0.76485	0.76874
LT24	0.86366	0.86873	0.89320	0.89590

N = 80

	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.58278	0.58964	0.70619	0.71121
IN32	0.66836	0.68810	0.70355	0.71433
IN33	0.93079	0.93632	0.95340	0.95629
LT21	0.90768	0.91183	0.93728	0.93949
LT22	0.84638	0.85811	0.86268	0.86873
LT23	0.71884	0.72314	0.80488	0.80756
LT24	0.89342	0.89680	0.92162	0.92341

The dependability of the assessments with two and four items, and with one and two raters, is reported in Table 17 for school means based on 20, 50, or 80 students. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a school gets correct. As can be seen, the assessments have a dependability coefficient in excess of .70 only for a limited number of conditions involving large school sizes. As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 17: DEPENDABILITY INDICES (SCHOOLS)—CONNECTICUT

	N = 20			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.18495	0.18773	0.29106	0.29448
IN32	0.30451	0.32130	0.35272	0.36373
IN33	0.67289	0.68458	0.77527	0.78297
LT21	0.60763	0.61513	0.72092	0.72617
LT22	0.51199	0.52950	0.57134	0.58207
LT23	0.54086	0.55071	0.62039	0.62682
LT24	0.65523	0.66255	0.73276	0.73732

	N = 50			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.20079	0.20208	0.32434	0.32603
IN32	0.46732	0.48281	0.54136	0.55162
IN33	0.72110	0.72642	0.82501	0.82848
LT21	0.65215	0.65558	0.77376	0.77617
LT22	0.65895	0.67037	0.73085	0.73782
LT23	0.64363	0.64916	0.73975	0.74338
LT24	0.72891	0.73251	0.81526	0.81751

	N = 80			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.20518	0.20602	0.33389	0.33501
IN32	0.53942	0.55220	0.62492	0.63341
IN33	0.73425	0.73769	0.83846	0.84069
LT21	0.66431	0.66653	0.78820	0.78976
LT22	0.70989	0.71813	0.78570	0.79071
LT23	0.67573	0.67952	0.77712	0.77962
LT24	0.74999	0.75238	0.83887	0.84036

The standard errors for relative and absolute decisions are reported in Tables 18 and 19, respectively. These are the standard errors that could be used to make confidence intervals around the schools' mean scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and would need to be transformed for use with scores other than the average item score. Despite the lower dependability/generalizability coefficients for the schools, the standard errors are lower than expected. The poorer generalizability coefficients are in part due to the differences in the universe score variance associated with schools versus students, and the lower standard errors still provide a useful way of examining school means.

TABLE 18: STANDARD ERRORS FOR RELATIVE DECISIONS (SCHOOLS)—CONNECTICUT

	N = 20			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.24996	0.24183	0.20704	0.20215
IN32	0.20650	0.19709	0.19148	0.18646
IN33	0.20909	0.19527	0.17935	0.17138
LT21	0.22950	0.21979	0.20077	0.19526
LT22	0.21172	0.20208	0.19827	0.19318
LT23	0.27658	0.27050	0.23898	0.23547
LT24	0.23896	0.23241	0.21711	0.21353

	N = 50			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.20286	0.19887	0.15883	0.15629
IN32	0.13159	0.12568	0.12164	0.11847
IN33	0.14779	0.14002	0.12265	0.11801
LT21	0.16803	0.16275	0.14038	0.13724
LT22	0.13390	0.12781	0.12540	0.12218
LT23	0.21630	0.21320	0.17589	0.17399
LT24	0.16993	0.16626	0.14790	0.14580

	N = 80			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.18926	0.18660	0.14428	0.14253
IN32	0.10481	0.10017	0.09658	0.09409
IN33	0.12795	0.12237	0.10374	0.10032
LT21	0.14874	0.14503	0.12065	0.11837
LT22	0.10586	0.10104	0.09914	0.09659
LT23	0.19839	0.19629	0.15619	0.15485
LT24	0.14772	0.14509	0.12473	0.12317

TABLE 19: STANDARD ERRORS FOR ABSOLUTE DECISIONS (SCHOOLS)—CONNECTICUT

	N = 20			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.46955	0.46527	0.34909	0.34621
IN32	0.22486	0.21624	0.20156	0.19679
IN33	0.32716	0.31851	0.25263	0.24704
LT21	0.37478	0.36891	0.29018	0.28640
LT22	0.24259	0.23423	0.21523	0.21055
LT23	0.29228	0.28653	0.24814	0.24477
LT24	0.31025	0.30524	0.25829	0.25529

	N = 50			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.44626	0.44446	0.32284	0.32159
IN32	0.15885	0.15399	0.13695	0.13414
IN33	0.29182	0.28796	0.21611	0.21350
LT21	0.34062	0.33805	0.25219	0.25046
LT22	0.17876	0.17424	0.15079	0.14812
LT23	0.23605	0.23321	0.18816	0.18638
LT24	0.26083	0.25846	0.20360	0.20208

	N = 80			
	2 Tasks		4 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
IN31	0.44024	0.43911	0.31593	0.31514
IN32	0.13749	0.13399	0.11527	0.11319
IN33	0.28229	0.27981	0.20596	0.20426
LT21	0.33153	0.32988	0.24177	0.24064
LT22	0.15885	0.15568	0.12977	0.12784
LT23	0.21975	0.21785	0.16989	0.16866
LT24	0.24694	0.24537	0.18745	0.18641

DELAWARE

The data for Delaware are based on a statewide assessment during spring 1995 (similar findings were found for two preceding years and are not included in this report). The three assessment programs are mathematics, reading (open-ended problems with a single rater), and writing (one problem with two raters) in grades 3, 5, 8, and 10. To examine the generalizability at the student and school levels, 10 schools were randomly drawn from a pool of schools with at least 50 students, and 50 students were randomly drawn within each school. The writing program is analyzed separately from the reading and mathematics because of the difference in the designs. The variance components for the reading and mathematics assessments are reported in Table 20. As can be seen, the largest source of error is the task-by-pupil interaction, and the variance component associated with pupils is relatively large.

TABLE 20: VARIANCE COMPONENTS—DELAWARE

Test*	Sc	Pu:Sc	Tsk	Tsk*Sc	Pu*Tsk:Sc,E
READ03	0.0023480	0.038700	0.013112	0.00066	0.076
READ05	0.0018138	0.028077	0.008344	0.00004	0.106
READ08	0.0000040	0.027900	0.018534	0.00000	0.086
READ10	0.0028780	0.043100	0.029880	0.00104	0.111
MATH03	0.0043929	0.034000	0.036504	0.00354	0.107
MATH05	0.0017073	0.026636	0.054370	0.00080	0.078
MATH08	0.0000000	0.065900	0.035454	0.00124	0.121
MATH10	0.0054560	0.062400	0.025896	0.00104	0.089

Sc=School; Pu=Pupil; Tsk=Task; E=Error *Test is given as content area and grade level.

Student-Level Analyses

The generalizability of the assessments with 5, 10, and 15 items are reported in Table 21 (10 items approximates the actual test lengths). These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, all assessments have a generalizability coefficient in excess of .70 under the current measurement conditions.

TABLE 21: GENERALIZABILTY COEFFICIENTS—DELAWARE

Test	5 Items	10 Items	15 Items
READ03	0.72806	0.84263	0.88928
READ05	0.58496	0.73814	0.80873
READ08	0.61866	0.76441	0.82955
READ10	0.67233	0.80406	0.86025
MATH03	0.63458	0.77645	0.83896
MATH05	0.64266	0.78246	0.84364
MATH08	0.72940	0.84353	0.88995
MATH10	0.79027	0.88285	0.91873

The dependability indices of the assessments with 5, 10, and 15 items are reported in Table 22. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a student knows. As can be seen, all assessments except MATH05 have a dependability coefficient in excess of .70 under the current measurement conditions. As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 22: DEPENDABILITY INDICES—DELAWARE

Test	5 Items	10 Items	15 Items
READ03	0.69570	0.82055	0.87275
READ05	0.56646	0.72324	0.79674
READ08	0.57168	0.72747	0.80016
READ10	0.61830	0.76413	0.82934
MATH03	0.56625	0.72307	0.79660
MATH05	0.51555	0.68035	0.76148
MATH08	0.67632	0.80691	0.86242
MATH10	0.74532	0.85408	0.89774

The standard errors for relative and absolute decisions are reported in Tables 23 and 24, respectively. These are the standard errors that could be used to make confidence intervals around the students' scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and would need to be transformed for use with scores other than the average item score.

TABLE 23: STANDARD ERRORS FOR RELATIVE DECISIONS—DELAWARE

Test	5 Items	10 Items	15 Items
READ03	0.12382	0.08756	0.071489
READ05	0.14563	0.10298	0.084079
READ08	0.13115	0.09274	0.075719
READ10	0.14969	0.10585	0.086425
MATH03	0.14869	0.10514	0.085845
MATH05	0.12554	0.08877	0.072480
MATH08	0.15636	0.11056	0.090274
MATH10	0.13419	0.09489	0.077477

TABLE 24: STANDARD ERRORS FOR ABSOLUTE DECISIONS—DELAWARE

Test	5 Items	10 Items	15 Items
READ03	0.13399	0.09475	0.07736
READ05	0.15125	0.10695	0.08732
READ08	0.14459	0.10224	0.08348
READ10	0.16848	0.11913	0.09727
MATH03	0.17149	0.12126	0.09901
MATH05	0.16320	0.11540	0.09422
MATH08	0.17759	0.12558	0.10253
MATH10	0.15227	0.10767	0.08792

Parallel analyses are presented for the writing assessments in Tables 25, 26, and 27. In the variance components, all effects are confounded with some term involving students, and consequently relative and absolute standard errors do not differ, leading to the same level of generalizability for either type of decisions. Best conditions can be seen with four raters on a single item.

TABLE 25: VARIANCE COMPONENTS—DELAWARE

Test	Sc	Pu:Sc	Rat:Pu:Sc,E
WRIT03	0.02446	0.2115	0.28098
WRIT05	0.00910	0.1745	0.28235
WRIT08	0.01620	0.2315	0.40354
WRIT10	0.04505	0.4850	0.58840

Sc=School; Pu=Pupil; Tsk=Task; Rat=Rater; E=Error

TABLE 26: GENERALIZABILITY COEFFICIENTS/DEPENDABILITY INDICES—DELAWARE

Test	1 Rater	2 Raters	4 Raters
WRIT03	0.45646	0.62680	0.77060
WRIT05	0.39403	0.56531	0.72230
WRIT08	0.38035	0.55109	0.71059
WRIT10	0.47391	0.64307	0.78276

TABLE 27: STANDARD ERRORS—DELAWARE

Test	1 Rater	2 Raters	4 Raters
WRIT03	0.53007	0.37482	0.26504
WRIT05	0.53137	0.37573	0.26568
WRIT08	0.63525	0.44919	0.31762
WRIT10	0.76707	0.54240	0.38354

School-Level Analyses

The generalizability of the mathematics and reading assessments with 5, 10, and 15 items are reported in Table 28 for examining school means based on 20, 50, and 80 students within each school. These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, both assessments have a generalizability coefficient of approximately .70 or above under the current measurement conditions when the school sizes are 50 or 80 for all of the grades except grade 8. The grade 8 results seem to be linked to the 10 schools selected randomly and their lack of variability (these low levels of dependability did not replicate in the two earlier years of data). However, generalizability does not exceed .70 with a school size of 20.

TABLE 28: GENERALIZABILITY COEFFICIENTS (SCHOOLS)—DELAWARE

	N = 20		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.45372	0.49651	0.51263
READ05	0.42323	0.48347	0.50756
READ08	0.00177	0.00219	0.00237
READ10	0.45316	0.50562	0.52592
MATH03	0.55812	0.62918	0.65707
MATH05	0.42906	0.48653	0.50927
MATH08	0.00000	0.00000	0.00000
MATH10	0.56399	0.59792	0.61015

	N = 50		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.65992	0.70299	0.71863
READ05	0.64610	0.69995	0.71996
READ08	0.00442	0.00545	0.00591
READ10	0.65528	0.70782	0.72726
MATH03	0.70751	0.77876	0.80580
MATH05	0.62953	0.68953	0.71215
MATH08	0.00000	0.00000	0.00000
MATH10	0.75069	0.78099	0.79164

	N = 80		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.74451	0.78456	0.79889
READ05	0.74405	0.78818	0.80408
READ08	0.00705	0.00869	0.00942
READ10	0.73752	0.78645	0.80423
MATH03	0.75826	0.82797	0.85414
MATH05	0.71278	0.76983	0.79093
MATH08	0.00000	0.00000	0.00000
MATH10	0.81842	0.84573	0.85524

The dependability of the assessments with 5, 10, and 15 items is reported in Table 29 for school means based on 20, 50, or 80 students. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a school gets correct. As can be seen, no assessment has a dependability coefficient in excess of .70 under any condition (again, the grade 8 results are anomalous, due to lack of variation among schools). As expected, generalizability is lower for the absolute interpretations than for the relative interpretations.

TABLE 29: DEPENDABILITY INDICES (SCHOOLS)—DELAWARE

	N = 20		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.30113	0.38873	0.43047
READ05	0.30462	0.39551	0.43919
READ08	0.00067	0.00109	0.00137
READ10	0.23347	0.33157	0.38557
MATH03	0.28954	0.41316	0.48172
MATH05	0.11494	0.19084	0.24470
MATH08	0.00000	0.00000	0.00000
MATH10	0.36733	0.46574	0.51142

	N = 50		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.37991	0.50482	0.56695
READ05	0.40522	0.52947	0.58974
READ08	0.00087	0.00155	0.00209
READ10	0.27758	0.40800	0.48375
MATH03	0.32516	0.47279	0.55711
MATH05	0.12566	0.21576	0.28351
MATH08	0.00000	0.00000	0.00000
MATH10	0.43833	0.56978	0.63306

	N = 80		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.40650	0.54555	0.61575
READ05	0.44169	0.57845	0.64502
READ08	0.00094	0.00173	0.00241
READ10	0.29135	0.43294	0.51664
MATH03	0.33548	0.49049	0.57979
MATH05	0.12866	0.22304	0.29521
MATH08	0.00000	0.00000	0.00000
MATH10	0.46059	0.60348	0.67309

The standard errors for relative and absolute decisions are reported in Tables 30 and 31, respectively. These are the standard errors that could be used to make confidence intervals around the schools' mean scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and would need to be transformed for use with scores other than the average item score. Despite the lower dependability/generalizability coefficients for the schools, the standard errors are lower than expected. The poorer generalizability coefficients are in part due to the differences in the universe score variance associated with schools versus students, and the lower standard errors still provide a useful way of examining school means.

TABLE 30: STANDARD ERRORS FOR RELATIVE DECISIONS (SCHOOLS)—DELAWARE

	N = 20		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.053170	0.048795	0.047248
READ05	0.049718	0.044021	0.041951
READ08	0.047487	0.042720	0.041008
READ10	0.058932	0.053047	0.050935
MATH03	0.058975	0.050882	0.047882
MATH05	0.047664	0.042448	0.040560
MATH08	0.068942	0.063435	0.061490
MATH10	0.064946	0.060572	0.059042

	N = 50		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.034785	0.031496	0.030321
READ05	0.031520	0.027884	0.026562
READ08	0.030033	0.027019	0.025936
READ10	0.038910	0.034467	0.032853
MATH03	0.042615	0.035327	0.032537
MATH05	0.031697	0.027726	0.026269
MATH08	0.045277	0.041037	0.039522
MATH10	0.042568	0.039115	0.037895

	N = 80		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.028386	0.025392	0.024312
READ05	0.024979	0.022079	0.021023
READ08	0.023743	0.021360	0.020504
READ10	0.032004	0.027955	0.026469
MATH03	0.037423	0.030212	0.027389
MATH05	0.026229	0.022593	0.021244
MATH08	0.037071	0.033151	0.031737
MATH10	0.034792	0.031548	0.030389

TABLE 31: STANDARD ERRORS FOR ABSOLUTE DECISIONS (SCHOOLS)—DELAWARE

	N = 20		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.07382	0.060763	0.055736
READ05	0.06435	0.052652	0.048126
READ08	0.07721	0.060650	0.054012
READ10	0.09721	0.076171	0.067722
MATH03	0.10382	0.078990	0.068748
MATH05	0.11466	0.085081	0.072594
MATH08	0.10883	0.087002	0.078387
MATH10	0.09694	0.079111	0.072197

	N = 50		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.06191	0.047992	0.042349
READ05	0.05160	0.040149	0.035522
READ08	0.06789	0.050827	0.043684
READ10	0.08654	0.064622	0.055420
MATH03	0.09548	0.069989	0.059095
MATH05	0.10899	0.078776	0.065687
MATH08	0.09561	0.072315	0.062655
MATH10	0.08361	0.064184	0.056235

	N = 80		
Test	**5 Tasks**	**10 Tasks**	**15 Tasks**
READ03	0.05855	0.044226	0.038278
READ05	0.04788	0.036357	0.031595
READ08	0.06535	0.048059	0.040694
READ10	0.08367	0.061396	0.051890
MATH03	0.09328	0.067551	0.056425
MATH05	0.10753	0.077120	0.063843
MATH08	0.09201	0.068150	0.058059
MATH10	0.07994	0.059874	0.051477

Reported in Table 32 are the school-level results for the writing assessment. As can be seen, dependability exceeds .70 with larger school sizes but not when school size is small.

TABLE 32: GENERALIZABILITY COEFFICIENTS/DEPENDABILITY INDICES (SCHOOL)—DELAWARE

	N = 20		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.49833	0.58156	0.63455
WRIT05	0.28489	0.36570	0.42614
WRIT08	0.33784	0.42785	0.49361
WRIT10	0.45634	0.53625	0.58770

	N = 50		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.71292	0.77651	0.81276
WRIT05	0.49898	0.59039	0.64992
WRIT08	0.56054	0.65151	0.70904
WRIT10	0.67726	0.74298	0.78087

	N = 80		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.79893	0.84754	0.87414
WRIT05	0.61442	0.69753	0.74813
WRIT08	0.67114	0.74945	0.79588
WRIT10	0.77051	0.82223	0.85078

The associated standard errors are presented in Table 33. They are lower than the associated student-level standard errors, pointing to the low school variance components as the reason for the lower dependability levels.

TABLE 33: STANDARD ERRORS (SCHOOL)—DELAWARE

	N = 20		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.15692	0.13266	0.11869
WRIT05	0.15114	0.12563	0.11070
WRIT08	0.17819	0.14719	0.12892
WRIT10	0.23167	0.19738	0.17778

	N = 50		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.09924	0.08390	0.07507
WRIT05	0.09559	0.07946	0.07001
WRIT08	0.11270	0.09309	0.08153
WRIT10	0.14652	0.12484	0.11244

	N = 80		
Test	**1 Rater**	**2 Raters**	**4 Raters**
WRIT03	0.07846	0.066331	0.059345
WRIT05	0.07557	0.062817	0.055350
WRIT08	0.08910	0.073593	0.064458
WRIT10	0.11583	0.098692	0.088889

NORTH CAROLINA

The data for North Carolina are based on a field test assessment during spring 1995.[3] (Note: Some of these assessments with the weaker results were dropped before assessments were created for statewide use.) The two assessment programs are (1) mathematics, reading, and social studies in grades 3 through 8 (open-ended problems with two raters) and (2) English at the high school level (one problem with two raters). To examine the generalizability at the student level, all students with complete data were included for math, reading, and social studies. Because only 10 percent of the field test responses were scored by two raters, there were approximately 40 to 60 students; therefore, school-level analyses were not possible. In contrast, the English assessment was administered in 22 schools, with 20 responses from each school scored by two raters. The English program is analyzed separately from the reading, social studies, and mathematics program because of the differences in the designs. The variance components for the reading, social studies, and mathematics assessments are reported in Table 34. As can be seen, the largest source of error was usually the task-by-pupil interaction (consistent with Brennan, 1996) or the rater effect, which included the higher-order interaction of rater-by-task-by-pupil. In addition, the variance component associated with pupils is relatively large.

[3] Similar findings were found for two other forms in all areas except English and are not included in this report.

TABLE 34: VARIANCE COMPONENTS—NORTH CAROLINA

Test	Pu	Tsk	Tsk*Pu	Rat:Tsk*Pu,E
MATH03	0.41220	0.07712	0.36213	0.29673
MATH04	0.20658	0.15744	0.55551	0.32599
MATH05	0.03850	0.06315	0.23790	0.19319
MATH06	0.07350	0.48445	0.34844	0.21911
MATH07	0.24908	0.32944	0.33966	0.48969
MATH08	0.18733	0.24988	0.02150	0.97799
SOCST4	0.33325	0.02085	0.02379	0.61943
SOCST5	0.35900	0.01733	0.06155	0.33691
SOCST6	0.07537	0.17172	0.05488	0.45625
SOCST7	0.13281	0.01890	0.00000	0.52853
SOCST8	0.14596	0.01846	0.00000	0.67729
READ03	0.17923	0.01982	0.00000	0.42821
READ04	0.02254	0.21703	0.00000	1.47265
READ05	0.15252	0.02584	0.00000	0.66681
READ06	0.18650	0.62089	0.19351	0.59999
READ07	0.19120	0.07574	0.13077	0.33346
READ08	0.19874	0.02734	0.00000	0.67861

Sc=School; Pu=Pupil; Tsk=Task; Rat=Rater; E=Error

Student-Level Analyses

The generalizability of the assessments with 5 or 10 items and scored by one or two raters is reported in Table 35 (5 tasks approximates the actual test lengths). These coefficients represent generalizability when the decisions being made from the assessments are relative or comparative. As can be seen, some of the assessments in the lower grade levels have a generalizability coefficient in excess of .70 under the current measurement conditions. However, the results are varied.

TABLE 35: GENERALIZABILITY COEFFICIENTS—NORTH CAROLINA

	5 Tasks		10 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
MATH03	0.75776	0.80148	0.86219	0.88980
MATH04	0.53955	0.58976	0.70092	0.74195
MATH05	0.30869	0.36528	0.47176	0.53509
MATH06	0.39302	0.44518	0.56428	0.61609
MATH07	0.60027	0.68059	0.75021	0.80994
MATH08	0.48377	0.64724	0.65209	0.78585
SOCST4	0.72149	0.83323	0.83821	0.90903
SOCST5	0.81834	0.88642	0.90010	0.93979
SOCST6	0.42441	0.57113	0.59591	0.72703
SOCST7	0.55682	0.71533	0.71533	0.83404
SOCST8	0.51867	0.68306	0.68306	0.81169
READ03	0.67667	0.80716	0.80716	0.89329
READ04	0.07107	0.13271	0.13271	0.23433
READ05	0.53350	0.69580	0.69580	0.82061
READ06	0.54027	0.65393	0.70153	0.79076
READ07	0.67313	0.76266	0.80464	0.86535
READ08	0.59421	0.74546	0.74546	0.85417

The dependability of the assessments with 5 or 10 items scored by one or two raters is reported in table 36. These coefficients represent generalizability when the decisions being made from the assessments are absolute and the interpretations are linked to the percentage of the content domain that a student knows. As can be seen, the results of the generalizability analyses were mixed. As expected, generalizability is

lower for the absolute interpretations than for the relative interpretations presented in Table 35.

TABLE 36: DEPENDABILITY INDICES—NORTH CAROLINA

	5 Tasks		10 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
MATH03	0.73687	0.77814	0.84850	0.87523
MATH04	0.49855	0.54112	0.66538	0.70224
MATH05	0.28031	0.32619	0.43788	0.49192
MATH06	0.25889	0.28055	0.41130	0.43817
MATH07	0.51802	0.57675	0.68249	0.73157
MATH08	0.42848	0.55194	0.59991	0.71129
SOCST4	0.71503	0.82463	0.83384	0.90389
SOCST5	0.81193	0.87890	0.89620	0.93555
SOCST6	0.35564	0.45319	0.52468	0.62372
SOCST7	0.54813	0.70105	0.70812	0.82426
SOCST8	0.51195	0.67146	0.67721	0.80344
READ03	0.66669	0.79300	0.80002	0.88455
READ04	0.06252	0.10570	0.11767	0.19118
READ05	0.52403	0.67977	0.68769	0.80936
READ06	0.39734	0.45557	0.56870	0.62597
READ07	0.63905	0.71921	0.77978	0.83667
READ08	0.58465	0.73047	0.73789	0.84425

The standard errors for relative and absolute decisions are reported in Tables 37 and 38, respectively. These are the standard errors that could be used to make confidence intervals around the students' scores. As expected, the standard errors are higher for the absolute interpretations than for the relative interpretations. The differences between the two assessments are again affected by scale and should be transformed for use with scores other than the average item score.

TABLE 37: STANDARD ERRORS FOR RELATIVE DECISIONS—NORTH CAROLINA

	5 Tasks		10 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
MATH03	0.36301	0.31953	0.25668	0.22594
MATH04	0.41988	0.37908	0.29690	0.26805
MATH05	0.29363	0.25865	0.20763	0.18289
MATH06	0.33691	0.30265	0.23823	0.21401
MATH07	0.40727	0.34191	0.28798	0.24176
MATH08	0.44710	0.31953	0.31615	0.22594
SOCST4	0.35867	0.25826	0.25362	0.18262
SOCST5	0.28230	0.21448	0.19961	0.15166
SOCST6	0.31973	0.23791	0.22608	0.16823
SOCST7	0.32512	0.22990	0.22990	0.16256
SOCST8	0.36805	0.26025	0.26025	0.18402
READ03	0.29265	0.20693	0.20693	0.14632
READ04	0.54271	0.38375	0.38375	0.27135
READ05	0.36519	0.25823	0.25823	0.18259
READ06	0.39837	0.31417	0.28169	0.22215
READ07	0.30471	0.24393	0.21546	0.17248
READ08	0.36841	0.26050	0.26050	0.18420

TABLE 38: STANDARD ERRORS FOR ABSOLUTE DECISIONS—NORTH CAROLINA

	5 Tasks		10 Tasks	
Test	1 Rater	2 Raters	1 Rater	2 Raters
MATH03	0.38366	0.34282	0.27129	0.24241
MATH04	0.45584	0.41855	0.32232	0.29596
MATH05	0.31440	0.28201	0.22232	0.19941
MATH06	0.45869	0.43415	0.32435	0.30699
MATH07	0.48141	0.42754	0.34041	0.30231
MATH08	0.49988	0.38997	0.35347	0.27575
SOCST4	0.36443	0.26621	0.25769	0.18824
SOCST5	0.28837	0.22241	0.20391	0.15727
SOCST6	0.36955	0.30157	0.26131	0.21324
SOCST7	0.33089	0.23798	0.23397	0.16828
SOCST8	0.37303	0.26725	0.26377	0.18897
READ03	0.29934	0.21630	0.21167	0.15295
READ04	0.58132	0.43666	0.41106	0.30876
READ05	0.37220	0.26805	0.26318	0.18954
READ06	0.53186	0.47210	0.37608	0.33382
READ07	0.32863	0.27322	0.23237	0.19320
READ08	0.37575	0.27079	0.26570	0.19148

Parallel analyses are presented for the high school English assessments in Tables 39 to 41. In the variance components, all effects are confounded with some term involving students, and consequently relative and absolute standard errors do not differ, leading to the same level of generalizability for either type of decision. As can be seen, the English assessment works well even with a single rater.

TABLE 39: VARIANCE COMPONENTS—NORTH CAROLINA

Test	Sc	Pu:Sc	Rat:Pu:Sc,E
ENGLISH	0.06605	0.5835	0.18298

Sc=School; Pu=Pupil; Tsk=Task; Rat=Rater; E=Error

TABLE 40: GENERALIZABILITY COEFFICIENTS/DEPENDABILITY INDICES—NORTH CAROLINA

Test	1 Rater	2 Raters	4 Raters
ENGLISH	0.78021	0.87654	0.93421

TABLE 41: STANDARD ERRORS—NORTH CAROLINA

Test	1 Rater	2 Raters	4 Raters
ENGLISH	0.42777	0.30248	0.21388

School-Level Analyses

The generalizability of the mathematics, social studies, and reading assessments could not be estimated because data were not collected with this intent and the distribution of students with two raters was limited to too few schools. For the English assessment, Table 42 presents the generalizability coefficients with one, two, or four raters for examining school means based on 20, 50, and 80 students within each school. These coefficients represent generalizability when the decisions being made from the assessments are absolute (because of confounding, these also represent the coefficients with relative decisions). As can be seen, the assessment has a generalizability coefficient in excess of .70 under the current measurement conditions when the school sizes are 50 or 80. However, generalizability does not exceed .70 with a school size of 20.

TABLE 42: GENERALIZABILITY COEFFICIENTS/DEPENDABILITY INDICES (SCHOOL)—NORTH CAROLINA

	N = 20		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.63282	0.66183	0.67735
	N = 50		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.81163	0.83030	0.83996
	N = 80		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.87332	0.88673	0.89359

The standard errors for absolute decisions are reported in Table 43. These are the standard errors that should be used to make confidence intervals around the schools' mean scores.

TABLE 43: STANDARD ERRORS (SCHOOL)—NORTH CAROLINA

	N = 20		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.19577	0.18371	0.17738
	N = 50		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.12381	0.11619	0.11218
	N = 80		
Test	**1 Rater**	**2 Raters**	**4 Raters**
ENGLISH	0.097883	0.091855	0.088688

Conclusions and Recommendations

The results of this study lead to several conclusions and recommendations.

- Most assessment programs are using PBAs that have a reasonable level of generalizability (above .70) at the student level. However, the level of generalizability is not as high as that found for more traditional forms of assessment. Thus, other advantages of PBAs—such as consequences for instruction—are necessary to justify their use.

- Results at the school level are mixed. While the standard errors are smaller with reasonable school sizes, allowing the use of school means with confidence intervals, the lower variance component sometimes associated with schools results in low values for the generalizability coefficients and the dependability indices. This may be a function of the sampling designs in Connecticut (pilot) and North Carolina (field test). Even the findings in grade 8 in Delaware, where a statewide assessment was used, may be a function of using only larger schools to estimate the variance components.

- The relative magnitude of the variance components and the largest sources of error are consistent with the findings of Brennan (1996), suggesting that number of tasks has a greater effect than number of raters.

- Results were better than those previously reported by Shavelson, Baxter, and Gao (1993). However, the lower number of tasks needed may be the result of multiple factors, including how narrowly the task domain is defined and the effects of high-stakes testing on teaching to the test. It should also be noted that Shavelson et al. based their conclusions on a minimum level of generalizability equal to .80. (Even at .80, these results suggest fewer tasks are needed.)

- Although limited data are available, it appears from the Connecticut assessments that tests with more extended responses need fewer items to achieve the same level of generalizability. However, even with extended responses, at least 2 tasks are needed to achieve reasonable levels of generalizability.

References

Brennan, R.L. (1996). Generalizability of performance assessments. In G.W. Phillips (Ed.), *Technical issues in large-scale performance assessment.* (NCES 96-802) Washington, DC: National Center for Education Statistics.

Brennan, R.L., Gao, X., & Colton, D.A. (1995). Generalizability analyses of Work Keys Listening and Writing Tests. *Educational and Psychological Measurement, 55,* 157–176.

Gamache, L.M., & Brennan, R.L. (1994, April). Issues of generalizability: Tasks, raters, and contexts for the NCBE-PT. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Gao, X., Brennan, R.L., & Shavelson, R.J. (1994, April). *Generalizability of group means for performance assessments under a matrix sampling design.* Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Shavelson, R.J., Baxter, G.P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30,* 215–232.

Appendix
Formulas Used in Generalizability Analyses

For all analyses, the universe score variance and the relative and absolute error variance formulae are given below. When referring to the tables, the generalizability coefficient is the universe score variance divided by the universe score variance plus the relative error variance. The dependability index is universe score variance divided by the universe score variance plus the absolute error variance. Finally, the standard errors are the square root of the error variances. For all equations below, the subscripts refer to school (s), pupil (p), task (t), rater (r), and random error (e).

Alabama – Student Level

(1) Universe Score Variance

$$\sigma^2_s + \sigma^2_{p:s}$$

(2) Relative Error Variance

$$\sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S,e}/n'_t$$

(3) Absolute Error Variance

$$\sigma^2_T/n'_t + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S,e}/n'_t$$

Alabama – School Level

(1) Universe Score Variance

$$\sigma^2_s$$

(2) Relative Error Variance

$$\sigma^2_{p:s}/n'_p + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S,e}/(n'_t n'_p)$$

(3) Absolute Error Variance

$$\sigma^2_T/n'_t + \sigma^2_{p:s}/n'_p + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S,e}/(n'_t n'_p)$$

Connecticut – Student Level

(1) Universe Score Variance

$$\sigma^2_s + \sigma^2_{p:s}$$

(2) Relative Error Variance

$$\sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S}/n'_t + \sigma^2_{R:PST,e}/(n'_r n'_t)$$

(3) Absolute Error Variance

$$\sigma^2_T/n'_t + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S}/n'_t + \sigma^2_{R:PST,e}/(n'_r n'_t)$$

Connecticut – School Level

(1) Universe Score Variance

$$\sigma^2_s$$

(2) Relative Error Variance

$$\sigma^2_{p:s}/n'_p + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S}/(n'_t n'_p) + \sigma^2_{R:PST,e}/(n'_r n'_p n'_t)$$

(3) Absolute Error Variance

$$\sigma^2_T/n'_t + \sigma^2_{p:s}/n'_p + \sigma^2_{T\bullet S}/n'_t + \sigma^2_{T\bullet P:S}/(n'_t n'_p) + \sigma^2_{R:PST,e}/(n'_r n'_p n'_t)$$

Delaware – Reading, Math

Same as Alabama

Delaware – Student Level (Writing)

(1) Universe Score Variance

$$\sigma^2_s + \sigma^2_{p:s}$$

(2) Relative and Absolute Error Variance

$$\sigma^2_{R:P:S,e}/n'_r$$

Delaware – School Level (Writing)

(1) Universe Score Variance

$$\sigma^2_s$$

(2) Relative and Absolute Error Variance

$$\sigma^2_{p:s}/n'_p + \sigma^2_{R:P:S,e}/(n'_r n'_p)$$

North Carolina – Student Level (Math, Social Studies, and Reading)

(1) Universe Score Variance

$$\sigma^2_p$$

(2) Relative Error Variance

$$\sigma^2_{T\bullet P}/n'_t + \sigma^2_{R:T\bullet P,e}/(n'_r n'_t)$$

(3) Absolute Error Variance

$$\sigma^2_T/n'_t + \sigma^2_{T\bullet P}/n'_t + \sigma^2_{R:T\bullet P,e}/(n'_r n'_t)$$

North Carolina – English

Same as Delaware Writing

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

X	This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.
	This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").